DOCUMENT RESUME

ED 482 280 TM 035 366

| | |
|---|---|
| AUTHOR | Landgraf, Kurt M. |
| TITLE | Using Assessments and Accountability To Raise Student Achievement. |
| INSTITUTION | Educational Testing Service, Princeton, NJ. |
| PUB DATE | 2003-00-00 |
| NOTE | 16p.; Based on the testimony of Kurt M. Landgraf before the Education Reform Subcommittee of the House Committee on Education and the Workforce on Measuring Success (Washington, DC, March 8, 2001). |
| PUB TYPE | Reports - Descriptive (141) |
| EDRS PRICE | EDRS Price MF01/PC01 Plus Postage. |
| DESCRIPTORS | *Academic Achievement; *Accountability; Achievement Tests; Educational Improvement; *National Competency Tests; *State Programs; *Test Use; Testing Problems; *Testing Programs |
| IDENTIFIERS | National Assessment of Educational Progress |

ABSTRACT

This document comments on the future of educational testing in the United States and the plans of the Bush administration for increased use of testing for educational accountability. The "achievement gap" does not appear to be closing. One of the keys to closing the gap is having the data to understand it so that teachers can use test results appropriately. The president's plan calls for school-by-school report cards with mathematics and reading tests broken down by ethnicity, gender, disability, and English proficiency. Sanctions and rewards based on closing achievement gaps and improving English proficiency can help, but creating an accountability system does not automatically produce a productive learning environment. The rewards/sanctions system needs to be planned carefully to avoid being trivial, counterproductive, or corrupted. President Bush's plans require testing some 22 million students in grades 3 through 8 each year in reading and mathematics. The plan also requires that such tests be aligned with the state's academic standards. To accomplish this, a major test creation and administration effort will be required in a number of states. This is doable given sufficient time and resources. Any testing program, however, is only as good as the weakest link in the process. The stakes are high, and it is essential that test developers implement safeguards established by the assessment profession. The president's testing program should go forward, but it should be done right. Recommendations are made to bring this about. (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 482 280

# Using Assessments and Accountability to Raise Student Achievement

TM035366

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

R. Coley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Based on the testimony
of Kurt M. Landgraf
President and Chief Executive Officer,
Educational Testing Service

ETS®
Educational
Testing Service



BEST COPY AVAILABLE

# Using Assessments and Accountability to Raise Student Achievement

*Based on the testimony of Kurt M. Landgraf, president and chief executive officer of Educational Testing Service, before the Education Reform Subcommittee of the House Committee on Education and the Workforce on Measuring Success (March 8, 2001)*

As the president of the nation's largest educational measurement institution, I understand the value of testing and the vital role it should play in education reform. Well-designed tests tied to standards and curriculum can provide useful information to guide instruction and help students learn. Test results can also provide useful data to guide sound education policy decisions. In the public schools, we need to provide resources and support to help teachers teach and help students learn and to monitor progress via well-designed assessments.

ETS®
*Educational Testing Service*

It calls for high standards, strong accountability, and annual standards-based assessments. Results from these tests will provide important information that the American people and policymakers need to move this nation forward and to ensure significant education reform. Most importantly, the president's plan targets the stubbornly persistent achievement gaps among different groups of students.

Without solid and frequent information gathered from student assessments, it will be difficult for us to know if each child is mastering the material appropriate for his or her age and grade. Yearly assessments will help provide teachers and school administrators with the critical information they need to enable each and every student to learn.

ETS supports the third- through eighth-grade testing plan, but testing alone is not enough. It is just one step in education reform. It is a misuse of tests when nothing is done to change poor results. If we take no action to improve teaching and learning, we will just be using children as "extras" in a high profile political drama while undermining the social and economic prospects of the nation in the process. In addition to giving tests, we must help teachers and students improve classroom achievement so that the results improve the next time we test.

*"I believe President Bush's education reform proposal, 'No Child Left Behind,' is the right thing for our country, and it is doable . . ."*

## Increasing Accountability in Closing the Achievement Gap

The "achievement gap"—the difference in school performance tied to race or ethnicity—does not appear to be closing. Data over a period of 30 years from the National Assessment of Educational Progress (NAEP) show that achievement among students overall has gradually increased in math and remained about the same in reading and science. But the gap between White and Black students has been widening over the past 10-15 years in mathematics and reading in middle and high school. The gap between Hispanic and non-Hispanic students also persists.

It is unconscionable that in the United States of America—which people from around the globe consider the "land of opportunity"—we have a test score achievement gap. There are many theories as to why it exists and what it will take to end it once and for all. One of the keys to closing the gap is having the data to understand it so that we can help teachers use test results appropriately; provide schools with well-targeted systems, tools, and resources; and hold schools accountable for eradicating the gap.

The president's plan calls for school-by-school report cards with mathematics and reading test results broken down by ethnicity, gender, poverty, disability, and English proficiency. These results—linked to school factors such as time on task in

Photographs by: Max Krupka

various subjects, teacher qualifications, preparation and placement, alignment of curriculum and standards, and instructional practices—will help educators diagnose problems and design remedies to improve student achievement across all groups.

The president got accountability right when he based his sanctions and rewards on closing achievement gaps and improving English proficiency. Like any good executive, he has focused attention on the areas where change must take place. Thoughtfully designing incentives and sanctions and targeting resources to identified needs—this is how we can make a difference.

Researchers have found that creating an accountability system does not automatically produce a productive learning environment. The rewards/sanctions system needs to be carefully planned if it is to avoid being trivial, counterproductive, or corrupted.

*"It is unconscionable that in the United States of America . . . we have a test score achievement gap."*

## Annual Testing in Reading and Math

Good testing, done right, is a good thing. Without standardized testing, parents and taxpayers can't know how much their students have learned relative to standards or to other students. Test results, used in conjunction with other information, help us make informed decisions about best practices in teaching. They can also help us compare our students' achievement with that of students in other nations. We often focus on "inputs" to education: how many books, how much money, how many teachers. These are very important. But the result—student learning—is what this enterprise is all about. If we are not measuring critical results accurately and often, we cannot know where we are going or how to get there.

The benefits of annually testing children as they develop foundational learning skills in grades three through eight are enormous. The key is to develop tests that measure the curriculum and for schools to use the

results to improve student learning. This means that scores must be published in a timely manner and that parents, teachers, and administrators must understand how to interpret the results. In addition, test results should lead to a plan of action to help students build toward mastery of standards. The testing exercise must become a learning event for students, teachers, and parents. Given sufficient resources from Washington and the states, this will happen.

Test results will help promote learning. The ultimate effect of clear standards, relevant curricula, well-trained educators, and valid assessments working in concert will be an upgraded education system, increased student achievement, closing of the achievement gap, and yes, assurance that no child is left behind. I agree with President Bush that there is no greater purpose than this.

## What It Will Take

President Bush's plan will require testing some 22 million students in grades three through eight every year in reading and math. That's 12 tests—one each in reading and math for each of the six grades—per state, or 600 tests per year. A recent study by the Education Commission of the States and press accounts report that 15 states already have tests in these subjects in those grades. But the president's plan requires—and rightly so—that such tests be aligned with the state's academic standards. Only seven of the 15 states currently use tests aligned with statewide academic standards for reading and math in all six grades. Eleven more states test in all but one of those grades, and three others test in all but two of those grades. But 21 states test in three or fewer of the six grades and would have to at least double the number of students they test annually. Thus, a major test creation and administration effort will be required of a number of states. This is an ambitious undertaking, but it is doable given sufficient time and resources.

The president's plan also calls for parents to get a report on how well their child is learning and for school-by-school report cards. Mathematics and reading results must be broken down by specified subgroups. Test

*Continued on pg 9*

# How Tests are Developed

Developing a high-quality test, even in just one subject for one grade, is a lengthy, multi-step undertaking. To ensure reliable data and prevent costly mistakes, we should spend what it takes to get it right the first time. Done properly, test development usually takes about 18-24 months, including refinements to the test form.

There are eight basic steps in the test development process:

1 **Defining purpose and objectives** – Careful consideration must be given to the students who will be taking the test and the purposes for which the test is being developed. This information will affect the content, the types of test questions, the length and difficulty of the test, and thus the time and cost.

2 **Convening development committees to write test specifications** – At ETS, our technical experts work with state officials and their designated experts on the subject standards to determine not only the content of the test but also the form it will take, the number and types of questions, and their level of difficulty. These specifications are based on a state's content standards and its initial statement of target performance levels.

3 **Question-writing and review** – Test questions are usually written by a combination of state-designated experts, testing company staff, teachers, and outside experts, depending on the state's requirements. Each question must be reviewed to ensure that it is clear and unambiguous, that reviewers agree on the intended correct response or the number of points to be given to responses to an open-ended question, that the question is fair to all test takers, and that it is in an appropriate editorial style.

4 **Pretesting** – To ensure fairness, reliability, and accuracy, pretesting is conducted before tests are administered on a large scale. Results of the pretest indicate the difficulty of each question and whether questions are ambiguous and therefore should be revised or discarded, or whether any answer choices should be revised or replaced.

5 **Data analysis, test assembly, and publication** – During this phase, test makers select questions that assess the required subject matter or skills. Both content and difficulty are considered in choosing items to match the requirements of the test specifications. After the test is assembled, other specialists, committee members, and outside experts ensure that the intended answer is the correct answer for each question and that the test specifications have been met.

6 **Test administration** – Standard testing procedures and security of testing materials are very important. Special accommodations are provided, according to specified guidelines, for students with disabilities. Make-up tests for absentees must also be planned for.

7 **Scoring** – Score ranges and cut points associated with proficiency levels are established based on the state's earlier specification of performance levels in conjunction with score data from a real test administration.

8 **Analysis and reporting** – Test specifications and questions may be readjusted or realigned.



BEST COPY AVAILABLE

*Continued from pg 7*

analysis results will determine the extent to which statistical specifications for difficulty, reliability, intercorrelations of subparts, etc. have been met. Discrepancies between desired and actual results will lead to improvements in the next form of the test.

Any testing program is only as good as the weakest link in this process. The proposed tests and the way results are used will demand greater validity, reliability, and measurement precision than ever before, particularly in view of their potential consequences for students and their schools.

The consequences of the proposed testing program are essential to accountability. When the stakes are high, however, it is important that test developers implement the safeguards established by the assessment profession. These include the following basic principles that protect students:

- Students should have adequate notice of the skills and content to be tested, along with other appropriate test-preparation material.
- Students should have access to the curriculum and to instruction that gives them the opportunity to learn the content and skills that are tested.
- Students should have equal access to any specific preparation for test taking.
- If the high stakes affect individual students, they should be given multiple opportunities to demonstrate their capabilities through repeated testing with alternate forms or through equivalent means.
- Scores from large-scale assessments should not be used alone if other information will increase the validity of the decision being made.

*"The testing exercise must become a learning event for students, teachers, and parents."*



9

## Cost Issues

The cost of developing tests is of paramount interest. The president's plan calls for the federal government to cover the development costs for new K-12 state tests. A number of factors will influence these costs. These include:

- **Type of test**—The test may be multiple-choice, open-ended, or a combination of both. It costs a few cents to score a multiple-choice answer sheet and from one to two dollars to score an essay.
- **Security issues**—A state may decide to use the same form of a test every year for five years or to use a new test every year for five years.
- **Administration procedures**—A state may decide that teachers should administer the test. Sometimes unannounced visits by proctors may be made to observe the test administration.

The greatest variable related to test cost is quality. Factors associated with the quality of a test include:

- Test design
- Development of the test questions, including who writes the questions, the procedures for review of the questions, and pilot testing
- Test forms, including the number of forms and how the forms are assembled and field tested
- Scoring accuracy involving multiple quality-control checks for electronic scoring, and rigorous training and quality control procedures for essay scoring
- Data analysis including multiple quality-control checks for data files and programs for analyses
- Reporting that should provide understandable and useful information to students, teachers, and parents

The stark reality of school and state budgets inevitably forces trade-offs. The availability of federal assistance will permit greater attention to quality and therefore improve the chances for valid, reliable, and useful results. The availability of federal funds might also be

*Continued on pg 12*

# Quality and Fairness in Testing

Ensuring the quality and fairness of tests is essential to test development. Factors to be considered include the characteristics of the questions on the test, the validity and reliability of the test, and the way in which the results are used.

- All tests and test questions should be subjected to thorough, professional reviews to eliminate symbols, words, phrases, art, and content that may be considered to have a gender or ethnic bias. Questions should reflect the multicultural nature of our society, with all groups represented with appropriate, positive references.

  In addition, statistical analysis should be used to identify specific questions on which minority-group test takers and majority test takers, matched according to their similar knowledge and/or skills on the subject tested, perform differently to a significant degree. Such questions should be reviewed by outside experts as to their fairness and removed if judged unfair.

- Reliability – Consistency throughout the test–and from one edition of the test to another–is a critical indicator of the accuracy of a test. Performance on one version of the test should reasonably predict performance on any other version of the test. If reliability is high, results will be similar, no matter which version a test taker completes or who scores an essay.

- Validity is the essential measure of whether the test is doing what it is supposed to do for a particular purpose. It is the extent to which inferences made and actions taken on the basis of the test scores are appropriate. Validity is based on logical and empirical evidence.

- The proper use of test scores is essential because the president's plan creates a testing landscape where test results will not just sit in a file folder. These results should be used to diagnose a student's needs, to help determine promotion to the next grade, or to suggest remediation. The test score data should inform subsequent action. This means that score data must be reported in time and in a format to serve these purposes.

  ETS is concerned that adding more volume to test score data, without the means to manage the data, will not inform instruction. Therefore, we suggest that Congress encourage states and districts to undertake the development of data management systems that will support serious analysis of the test results by the professionals responsible for advancing student achievement. Specifically, we recommend that Congress authorize and fund a challenge grant program to utilize technology in the service of test administration as well as the management of assessment data.



BEST COPY AVAILABLE

*Continued from pg 10*

used to harness the tremendous power of technology in the delivery and management of school assessments.

The cost of testing can be lowered through economies of scale, scope, and experience. The more students tested, the lower the cost. The per-student cost is expected to decline as fixed costs (e.g., for test development, distribution, and test preparation and scoring) are divided by a larger test population. When the same test administration is used for several purposes, such as to test the same students in more than one subject, the cost of tests per subject declines as more subjects are included. Testing costs may also decline as simpler and less expensive processes are discovered.

## How to Do It Right

I believe the president's testing plan should go forward, but it should be done right and it should be done well. In order to do it right, I recommend the following:

- Continued development of unambiguous standards in each state that the education community and the public accept as meaningful
- State curricula that are linked to state standards
- Instructional materials that are linked to the curricula
- Professional development for teachers and administrators to understand the standards, know the curriculum, and skillfully use the learning materials
- The opportunity for all students to learn the curriculum's material
- Prior notice to students of testing requirements
- Assessments linked to the standards
- Alternative assessments for students with disabilities and those students who are nonnative speakers of English
- Effective remedial programs for students who fail, and a policy of nonretention if remediation is no better than promotion
- Communication with the public to enlist its support and understanding
- Resources to support the whole learning enterprise, not just the tests

The president's plan allows states three years to develop and implement the assessments. For some states, this will be insufficient time to do all that needs to be done. Taking more time to expand the range of experts and stakeholders involved in the process can make the difference between success and failure. We should balance the needed pressure for change with the needed time for doing it right.

Recent history tells us that developing standards and creating new tests aligned with those standards is a time-consuming process. The fresh evidence of states' recent experiences in implementing the testing requirements of Title I, mandated in 1994, is instructive. Only about 10 percent were able to comply in six years' time. Of the 34 states whose testing systems the Education Department has now evaluated, only 17 have received full approval for meeting the Title I requirements. Fourteen states have been granted extra time, and three states must agree to make changes by a specified deadline. The testing systems of 16 other states, the District of Columbia, and Puerto Rico are still under review, with decisions expected this spring.

*"The president's testing plan should go forward, but it should be done right and it should be done well."*

## The Use of NAEP In the President's Plan

The National Assessment of Educational Progress (NAEP), often called the "Nation's Report Card," is the most widely respected, nationally representative continuing assessment of what America's students know and can do in various subject areas. NAEP provides a comprehensive measure of students' learning at critical junctures in their school experience. ETS is extremely proud to have served as the prime contractor for NAEP since 1983.

13

This assessment has been conducted regularly since 1969. Until 1990, NAEP was solely a national assessment. Because the national NAEP samples were not, and are not currently, designed to support the reporting of accurate and representative state-level results, in 1988 Congress authorized a voluntary Trial State Assessment (TSA). Separate representative samples of students are selected for each jurisdiction that agrees to participate in TSA, and these jurisdictions receive reliable state-level data concerning the achievement of their students. In 1996, "Trial" was dropped from the title, based on numerous evaluations of the TSA program.

President Bush has proposed verifying state test scores by "confirming" them with NAEP results. For that to happen, all states would participate in the National Assessment, and NAEP fourth- and eighth-grade reading and math tests would be given every year instead of every two to four years. The meaning of "confirm" is operationally a complicated matter that will have to be considered by groups of experts in the coming months and must take into account the relationship between the contents of the NAEP assessments and state assessments.

Because NAEP is a congressionally mandated and widely respected broad survey of student achievement in the U.S., it is reasonable for the president to propose using NAEP as part of his plan. NAEP is a broad measure of content and skills and therefore provides invaluable information on what our children know and can do. However, how best to use NAEP in a confirmatory role deserves serious consideration.

As occurred with the TSA, I would suggest that the use of NAEP in its new proposed confirmatory role be conducted on a trial basis until such time as an independent evaluation certifies the rigor of the confirmations and the fairness of the process.

Most recently, 40 states have participated in NAEP, although 48 had signed up initially. Thus, the president's proposal that all states participate in NAEP's annual reading and math assessments, and that Congress fund administration of those tests, seems doable.

## Recommendations

Used properly, assessment can be the linchpin of an education reform strategy that spurs learning while monitoring results. For assessment to work effectively as a catalyst for reform, we should:

- Balance the needed pressure for change with the needed time for doing it right.
- Ensure proper safeguards for test scores used in high-stakes situations.
- Use NAEP as the instrument for confirming state assessment results, after additional study.
- Provide technical assistance, including that offered by comprehensive regional assistance centers, to help schools, districts, and states implement the president's plan prior to the imposition of consequences.
- Create a program of ongoing research to document the progress and outcomes of the "No Child Left Behind" plan. We need to know whether students as a whole and among various subgroups did learn more, whether the achievement gap was closed, what factors increased those outcomes, and at what cost.
- Urge Congress to include in its Elementary and Secondary Education Act reauthorization bill a new 21st Century State Assessment Challenge Grant program to support collaborative efforts by groups of states to develop prototypes for the electronic delivery of state assessments. Such a program will help move existing state-of-the-art assessment technologies into state K-12 systems, expediting the provision of assessment results to students, parents, teachers, administrators, and policymakers. Appropriate interventions could thus be applied sooner and more effectively to help assure that no child is left behind.

If implemented properly, President Bush's education reform plan will advance learning throughout the country. The education of all our children is the nation's top priority, and ETS wholeheartedly endorses this goal.



15

03748-008086 • y41M1.5 • Printed in the U.S.A.
I.N. 991364

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

# NOTICE

## Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").